Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants and mutations

Authors: Timothy Gilpatrick[1], Isac Lee[1], James E. Graham[2], Etienne Raimondeau[2], Rebecca Bowen[2], Andrew Heron[2], Fritz J Sedlazeck[3], Winston Timp[1]

1: Department of Biomedical Engineering, Johns Hopkins University (Baltimore, USA)
2: Oxford Nanopore Technologies (Oxford, UK)
3: Human Genome Sequencing Center, Baylor College of Medicine (Houston, USA)

Abstract

Nanopore sequencing technology offers a significant advancement through its ability to rapidly and directly interrogate native DNA molecules. Often we are interested only in interrogating specific areas at high depth, but this has proved challenging for long read sequencing with conventional enrichment methods[1]. Existing strategies are currently limited by high input DNA requirements, low yield, short (<5kb) reads, time-intensive protocols, and/or amplification or cloning (losing base modification information). In this paper, we describe a technique utilizing the ability of Cas9 to introduce cuts at specific locations and ligating nanopore sequencing adaptors directly to those sites, a method we term 'nanopore Cas9 Targeted-Sequencing' (nCATS).

We have demonstrated the ability of this method to generate median 165X coverage at 10 genomic loci with a median length of 18kb from a single flow cell, which represents a several hundred fold improvement over the 2-3X coverage achieved without enrichment. Using a panel of guide RNAs, we show that the high coverage data from this method enables us to (1) profile DNA methylation patterns at cancer driver genes, (2) detect structural variations at known hot spots, and (3) survey for the presence of single nucleotide mutations. Together, this provides a low-cost method that can be applied even in low resource settings to directly examine cellular DNA. This technique has extensive clinical applications for assessing medically relevant genes and has the versatility to be a rapid and comprehensive diagnostic tool. We demonstrate applications of this technique by examining the well characterized GM12878 cell line as well as three breast cell lines (MCF-10A, MCF-7, MDA-MB-231) with varying tumorigenic potential as a model for cancer.

Contributions

TG and WT constructed the study. TG performed the experiments. TG, IL and FS analyzed the data. JG, ER, RB, AH and TG developed the method. TG and WT wrote the paper

Introduction

Nanopore sequencing operates by reading the DNA base sequence through fluctuations in ionic current as a DNA molecule threads through a protein pore embedded in a membrane. The flagship product by ONT is the minION flow cell, which now generates 10+ gigabases worth of sequencing data, roughly 3X+ coverage of the human genome. Genomic characterization often calls for higher depths of sequencing, but limited to specific regions of the genome. There is a need for inexpensive but highly comprehensive tests that can scale to the demands of diagnostic laboratories to enable fast diagnostics at clinically informative loci. However, capture assays are still logistically challenging with long-read sequencing. Existing capture methods adapted to long read sequencing include CATCH-seq (dual cuts with Cas9 followed by size selection)[2], ligation to cuts with a less permissive endonuclease[3], cloning the region into an expression plasmid[4], region amplification[5], and long fragment hybridization capture methods[1,6]. Previous methods are limited by loss of native modifications, limited read length, requiring high input, low yield, or long protocols (>18hrs). We describe here an enrichment strategy using targeted cleavage with Cas9 for ligating nanopore sequencing adaptors (nanopore Cas9-Targeted Sequencing, or 'nCATS'), and show the use of this method for simultaneously assessing single nucleotide variants (SNVs), structural variants (SVs) and CpG methylation. The method can be completed on a benchtop in several hours, needs only ~3ug of genomic DNA and can be easily adapted to target a large number of loci in a single reaction. The potential of this technique, as combined with the low capital investment required for nanopore sequencing instruments (~$1K), could put a targeted sequencing assay in the hands of every pathology department, to evaluate cellular DNA for CpG methylation, structural rearrangements, and survey for nucleotide mutations.

Results

We enrich by selective ligation of nanopore sequencing adaptors at fresh cut sites created by active Cas9. To achieve enrichment, all preexisting DNA ends are dephosphorylated prior to the addition of the Cas9/guide RNA ribonucleoprotein complex (RNP). Cuts introduced by the RNPs thereby represent the majority of DNA ends with a 5-phosphate group and as a result nanopore sequencing adaptors are preferentially ligated to the DNA ends made by Cas9 cleavage (**Figure 1A**). By sequencing the native DNA strands, we avoid the need for PCR amplification and thereby maintain any modified nucleotides, which are distinguishable in nanopore electrical data[7]. Using human breast cell lines, we demonstrate applications of this method for characterizing breast cancer, focusing on genes where transcription and methylation have prognostic implications as well as sites believed to harbor large (>1kb) chromosomal deletions and finally single nucleotide variants (SNVs). We used the well-characterized GM12878 cell line to validate our ability to examine these features with the deeper coverage data from the nCATS method.

At each site evaluated, we used Cas9 to introduce two cuts flanking the region of interest. For sites without structural variations the guide RNAs (gRNAs) were placed 20-30kb apart, and for sites with deletions the gRNAs were designed to flank breakpoints by ~5kb. Using a single MinION flow cell for each sample, we targeted 10 loci (**Supplementary Table 1**) in three breast cell lines (MCF-10A, MCF-7, and MDA-MB-231). We performed parallel studies in the well-characterized GM12878 lymphoblast cell line, as well as a separate sequencing run for evaluating large annotated deletions in GM12878. The regions of interest for DNA methylation and structural variants were selected based on previous whole-genome nanopore data from our lab[8] as well as existing expression data in these breast cell lines[9]. To select point mutations we used existing deep coverage Illumina sequencing data of the the GM12878 cell line[10] as well as mutations in the MDA-MB-231 cell line found in the COSMIC database[11].

Yield and Coverage

Starting with ~3μg of genomic DNA from each cell line, we achieved significant enrichment (10 to 300-fold) at all of the regions evaluated from a single flowcell, with coverage ranging from 20X to 800X at each site. The total yield per flow cell ranged from 70K reads to 230K reads, and "on-target" fraction (fraction of reads that aligned to one of the targeted regions) ranged from 1.7% - 7.6%. Genome wide coverage analysis found the off-target reads to be distributed randomly across the genome, indicating they result from ligation of nanopore adaptors to random breakage points, with no clear evidence of off-target cleavage by Cas9. For example, in the GM12878 cell line, after quality filtering alignments (MAPQ > 30) there were only 2 genomic DNA sites sites outside target regions where coverage reached 25X, both at repetitive peri-centromeric sites and containing reads with lower mapping quality (MAPQ 30-50), suggesting the increased coverage to be the result of alignment errors in these poorly mappable

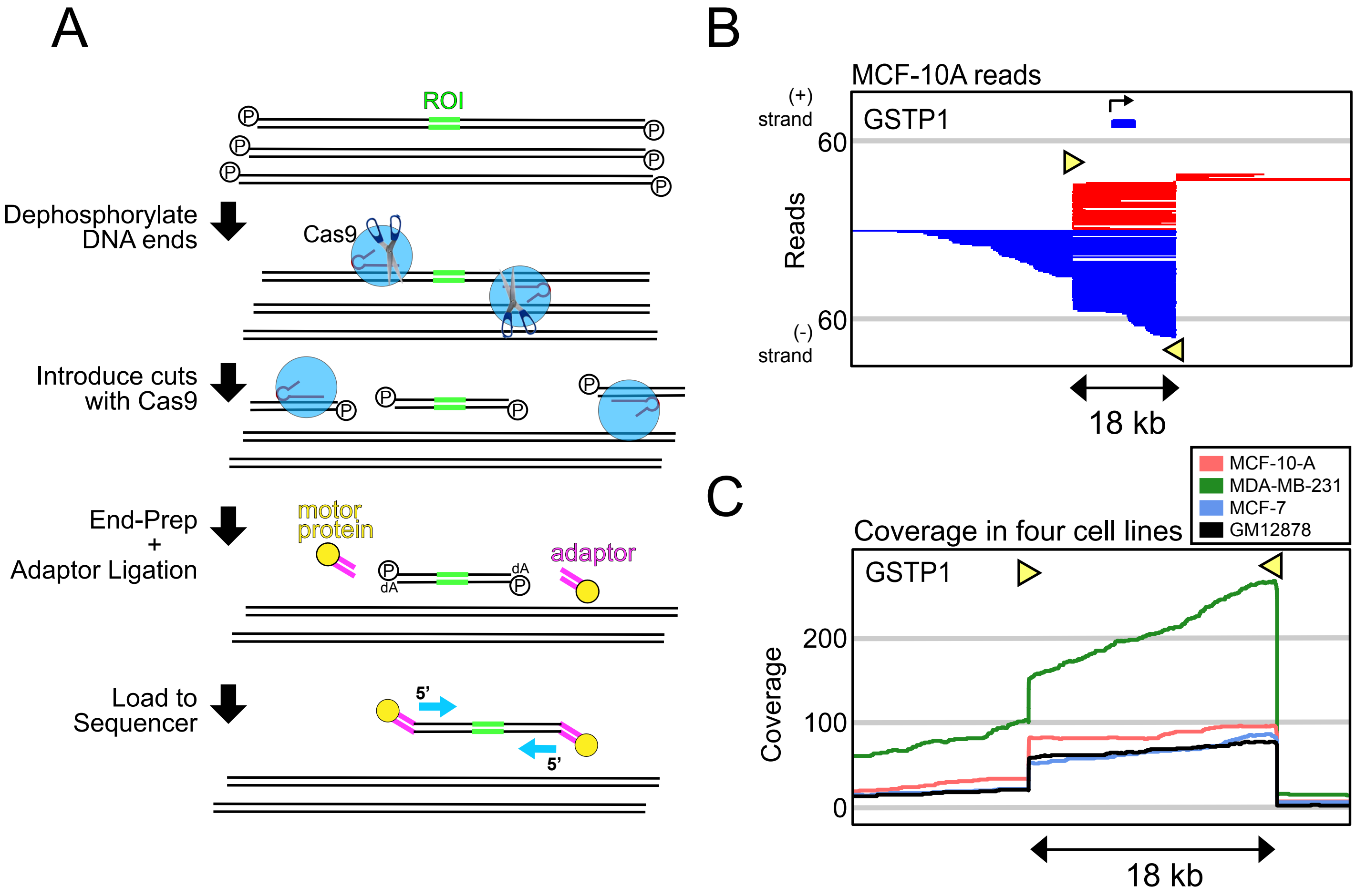


Figure 1: Method schematic and coverage data (A) Schematic of Cas9 enrichment operation. ROI = region of interest. DNA ends are first dephosphorylated, new cuts introduced with Cas9/guideRNA complex, nanopore sequencing adaptors are ligated to cuts around the ROI and and the sample is loaded to the nanopore sequencer. (B) Representative read-plots for the MCF-10-A cell line at GSTP1. Yellow triangles show Cas9 cut site and guideRNA direction (C) Coverage plots at the GSTP1 gene in the three breast cell lines and GM12878 cell line

regions. In contrast, all on-target alignments had MAPQ scores >50.

Figure 1B shows reads from the MCF-10-A cell line aligning to a representative locus (*GSTP1)*. Cas9 remains bound to the upstream (5') side of the gRNA after DNA cleavage, resulting in preferential ligation of adaptors onto the 3' side of the cut. A comparison of coverage between the four cell lines at the *GSTP1* locus is shown in **Figure 1C**. There was a difference in yield between the different guide RNAs and between cell lines, summarized in **Table 1**. This difference in coverage can be partially explained by both the stochastic formation of ribonucleoprotein complexes after combining the Cas9 with the gRNAs, as well as the different performance expected from the different gRNAs due to varying off-target mismatches and on-target binding performance. Further, a disparate number of reads were noted between cell lines which can be attributed to both (1) efforts made to keep the DNA as intact as possible, as the long strands of DNA may make it harder to get a uniform quantification; as well as (2) the aneuploidy which is known to exist in these immortalized breast cell lines. MDA-MB-231 and MCF-7 are near triploid and hypotetraploid, with modal chromosomal numbers of 64 and 82, respectively[12,13].

Methylation Studies

Five of our selected loci are in the promoters of genes where expression level and promoter methylation have prognostic implications in human cancer. Sites for methylation studies were selected by examining whole-genome nanopore methylation data from these breast cell lines[8], searching for differentially methylated promoters between the non-tumorigenic line MCF-10A and the tumorigenic lines MCF-7 and MDA-MB-231. Candidate loci were further filtered for genes with methylation status described as clinically informative in the literature, and examining existing RNA-seq expression data[9] to increase confidence that methylation is playing a regulatory role. To validate this method for studying DNA methylation, we compared nanopore methylation calls at these five loci with existing whole

cell line		GM12878		MCF-10A		MDA-MB-231		MCF-7	
total on-target reads		1375		4434		4134		1542	
LOCUS	region size (kb)	Coverage	(%) of on-target	Coverage	(%) of on-target	Coverage	(%) of on-target	Coverage	(%) of on-target
GPX1	17.8	130	9.45%	650	14.66%	400	9.68%	110	7.13%
GSTP1	18.2	75	5.45%	150	3.38%	250	6.05%	75	4.86%
KRT19	19.6	40	2.91%	60	1.35%	240	5.81%	90	5.84%
SLC12A4	13.6	180	13.09%	600	13.53%	400	9.68%	300	19.46%
TPM2	24.3	100	7.27%	300	6.77%	200	4.84%	60	3.89%
chr5 deletion	18.7	20	1.45%	70	1.58%	150	3.63%	50	3.24%
chr7 deletion	20.0	75	5.45%	350	7.89%	300	7.26%	100	6.49%
BRAF	12.3	50	3.64%	220	4.96%	200	4.84%	50	3.24%
KRAS	16.7	100	7.27%	200	4.51%	250	6.05%	100	6.49%
TP53	16.1	240	17.45%	850	19.17%	450	10.89%	180	11.67%

Table 1: Average coverage and percent of total on-target reads at ten loci in four cell lines

genome bisulfite sequencing (WGBS) data in GM12878[14]. Using line plots to show smoothed (loess) methylation, we compare methylation patterns at each of these regions. Specifically, we have plotted methylation data for two genes where CpG methylation is known to inform outcomes in human breast cancer (*GSTP1* and *KRT19)* (**Figure 2A**). Three additional genes (*GPX1*, *SLC12A4*, and *TPM2)* are included in **Supplementary Figure S1**. We note very similar methylation patterns between targeted nanopore and whole genome bisulfite data over all regions analyzed in GM12878, even with the noisy signal observed for many CpG sites. The aggregate correlation between the nanopore methylation calls and illumina methylation across all five regions was 0.84 (Pearson). Dot plots comparing methylation calls at each of the five genes are included in the supplemental materials (**Supplementary Fig. S2**), wherein we observe per-CpG methylation largely clustered at points reflecting completely methylated or completely unmethylated sites. We also note that there is a substantial reduction in the noise of the methylation signature around transcriptional start sites, suggesting that CpGs which play the largest regulatory role may have less inter-cellular variation in methylation status.

Because nanopore sequencing directly evaluates methylation patterns on native DNA strands, we are able to observe long-range methylation information on each DNA molecule. The read-level methylation plots demonstrate this phased methylation information (**Figure 2B; Supplementary Figure S1**), where each line directly represents a strand of cellular DNA and shows the methylation calls for evaluated CpGs. We use this to compare promoter methylation of the metabolism gene, glutathione S-transferase pi 1 (*GSTP1*), between breast cell lines. *GSTP1* is known to demonstrate promoter hypermethylation and transcriptional down-regulation in aggressive ER-positive breast cancer[15]. Correspondingly, we observe a dramatic increase in methylation at the *GTSP1* promoter in MCF-7: the only cell ER(+) breast cancer cell we examined (**Figure 2B**). Another example is the keratin family member gene: *KRT19*. *KRT19* expression is normally restricted to developing epithelial layers and not in mature mammary tissue[16]. *KRT19* is known to be expressed in breast cancers with poor prognostic outlook[17], and detection of *KRT19* has been used to demonstrate metastasis of breast cancer to lymph nodes and bone marrow[16]. We observe that the *KRT19* gene remains largely methylated in the non-tumorigenic MCF-10-A cell line, but the gene promoter becomes hypomethylated in both of the transformed cell lines, MCF-7 and MDA-MB-231 (**Figure 2B**). In **Figure 2C** and **Supplemental Figure S2**, we compare gene expression levels using existing RNA-seq data[9] from these breast cell lines. This demonstrates that mRNA levels of these genes follow the canonical inverse correlation between gene activity and promoter methylation, supporting the notion that CpG methylation is indeed playing a regulatory role.

Structural Variation

To validate the nCATS method for calling structural variations, we used available 10X genomics data from the Genome In a Bottle (GIAB) Consortium project[18] to identify large deletions present in the GM12878 cell line. We selected three heterozygous deletions, two with sizes of ~70kb and one ~150kb. Guide RNAs were designed to flank the deletion breakpoints by 5kb, resulting in reads of ~10 kb on the deleted allele, and spanning the region between cut sites on the non-deleted allele. Existing familial trio sequencing data on GM12878[10] annotates all heterozygous variants to their parental allele of

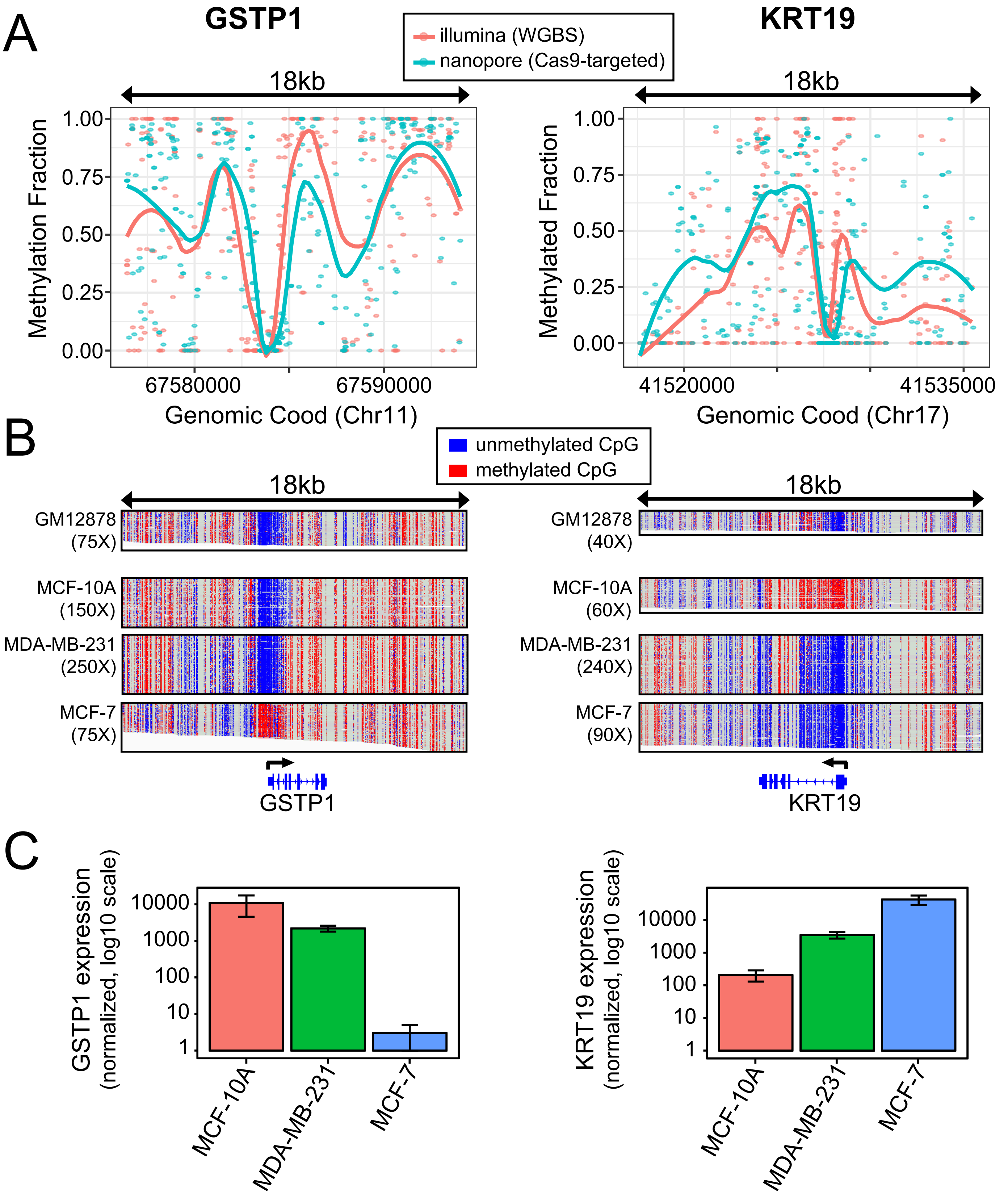


Figure 2: Methylation Studies and Comparison with Expression Data Data shown for *GSTP1* & *KRT19*; 3 additional genes in Supplemental Figure S1. (A) Methylation comparison between existing WGBS in GM12878 (red) (GEO: GSE86765) versus methylation calls from Cas9-targeted nanopore data in GM12878 (blue). Lines generated using loess smoothing (B) Read-level methylation plots using IGV showing methylation calls within nanopore reads around the gene body (C) Normalized expression levels from existing data (GEO: GSE75168) showing the inverse correlation between gene expression level and promoter methylation.

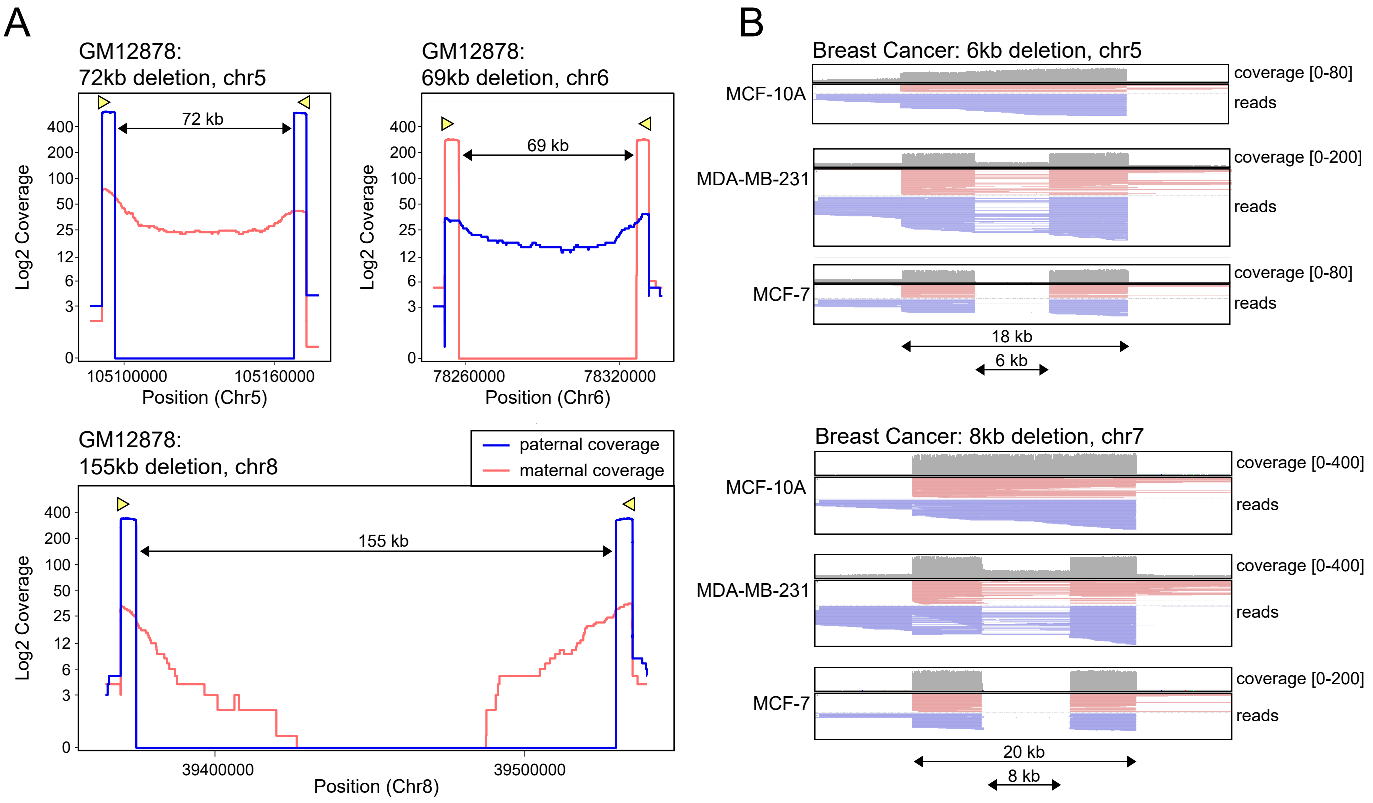


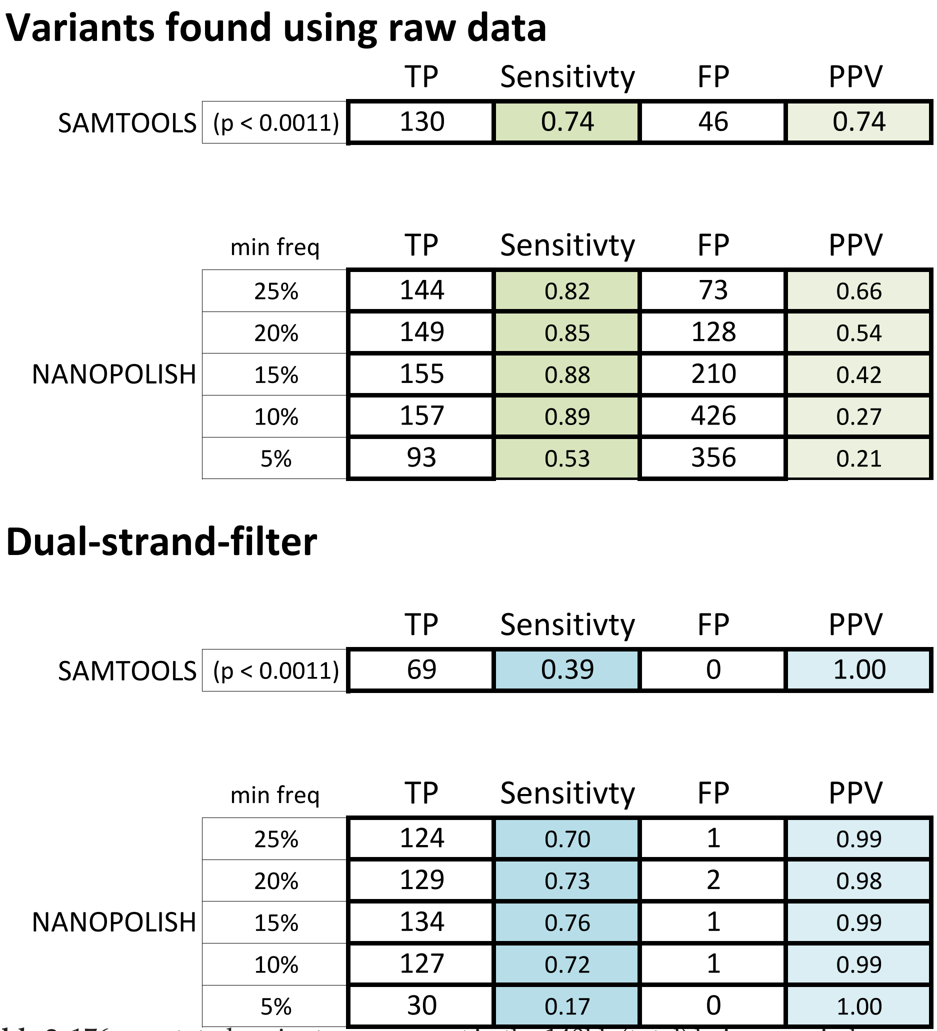
Figure 3: Structural Variation Studies (A) Coverage data from three deletions in the GM12878 lymphoblast cell line, showing increased coverage from the deleted (shorter) allele. Reads were segregated into parental allele of origin using the haplotype-aware tool HapCUT217 with reference data from familial trio sequencing9. Yellow triangles show Cas9 cut site and guideRNA direction (B) IGV plots showing reads and coverage at two deletions present in the MDA-MB-231 and MCF-7 cell line and absent in the MCF-10A cell line.

origin. Using this information, HapCUT2[19] was used to phase the reads and compare read lengths and read counts achieved from each allele. Interestingly, we found that the allele containing the deletion, with the correspondingly shorter distance between the cut sites, demonstrated a dramatically higher total number of reads (**Figure 3A; Supplementary Figure S3**). This may reflect a size bias introduced during purification, DNA fragmentation or library preparation. To confirm that this was a size-bias, we performed similar parental-allele segregation on sites without SVs and did not observe bias towards either parental allele (**Supplementary Figure S4**). The alignment data for these reads was then passed to the Sniffles variant caller[20] which identified all 3 of the deletions within 50nt from the annotated breakpoints in existing GIAB data (**Supplementary Table S2**). The imbalance of reads from the two alleles caused the software to initially identify these deletions as homozygous. We adjusted Sniffles parameters to call SVs as heterozygous if an allele was supported by even 0.1% of reads (see methods).

Using candidate deletions from whole genome nanopore sequencing data previously published from our lab[8], we selected two deletions present in the MDA-MB-231 and MCF-7 breast cancer lines, and absent in the MCF10A cell line. Plotting reads around these loci in IGV showed that coverage supported both of these reads as heterozygous deletions in MDA-MB-231 and homozygous deletions in MCF-7 (**Figure 3B**). We passed this data to the SV caller Sniffles, which identified the suspected ~7kb deletions. As suggested by the IGV plot, Sniffles called both of the the deletions as heterozygous in MDA-MB-231 and homozygous in the MCF-7 cell line (**Supplementary Table S3**). We also performed methylation studies on these regions, but did not note any difference in methylation patterns between the deleted and not-deleted allele (**Supplementary Figure S5**). This demonstrates how the high coverage nanopore data achieved through the nCATS method enhances our ability to evaluate and examine structural variants, and can be combined with methylation calls to study CpG methylation patterns at deletion breakpoints.

Single Nucleotide Variant Detection

Nanopore sequencing still has intrinsically high error rates due to the inability of the basecaller to distinguish between some k-mers and the difficulty in discriminating signal events in repetitive regions (e.g. homopolymers). We explored how the increased coverage would affect the ability to call variants using the variant-caller module of nanopolish. For initial validation, we once again turned to the GM12878 cell line, using the well-annotated platinum genome dataset as ground truth for SNVs[10]. We first examined the data at the *TP53* locus, where we had high read count (220X coverage) and balanced strand coverage. In the 16kb between gRNAs, there are 18 annotated SNVs in the platinum genome dataset for GM12878. We compared the variant-calling performance of the samtools package[21], which uses exclusively alignment data, with the nanopolish variant caller[7], which uses alignment data as well as raw nanopore electrical data. Using only the samtools package, 11 out of 18 SNVs were detected (Sensitivity: 0.61), with 10 false positives (Positive Predictive Value: 0.52). Positive Predictive Value (PPV) is defined as the fraction of true positive results out of all positive results. Using the default settings of nanopolish[7], the number of true positive variants called increased to 14 (sensitivity: 0.78), but the number of false positives also increased to 18 (PPV: 0.44).



Variants found using raw data

		TP	Sensitivty	FP	PPV
SAMTOOLS	(p < 0.0011)	130	0.74	46	0.74

	min freq	TP	Sensitivty	FP	PPV
NANOPOLISH	25%	144	0.82	73	0.66
	20%	149	0.85	128	0.54
	15%	155	0.88	210	0.42
	10%	157	0.89	426	0.27
	5%	93	0.53	356	0.21

Dual-strand-filter

		TP	Sensitivty	FP	PPV
SAMTOOLS	(p < 0.0011)	69	0.39	0	1.00

	min freq	TP	Sensitivty	FP	PPV
NANOPOLISH	25%	124	0.70	1	0.99
	20%	129	0.73	2	0.98
	15%	134	0.76	1	0.99
	10%	127	0.72	1	0.99
	5%	30	0.17	0	1.00

Table 2: 176 annotated variants are present in the 140kb (total) being queried across all sites. Comparing variant calling performance with samtools and nanopolish software tools, both on raw data as well as requiring variants to be supported by data from both strands (dual-strand-filter). Sensitivity = True positives result / All true positives Positive Predictive Value = True positive result / All positive results

The high number of false positive variants reported from nanopore data makes it difficult to discern true variants from the result of sequencing errors. On closer examination, we noted many false positives to result from errors on only one strand (**Supplementary Figure S6a**). Presumably, the basecaller was having systematic issues with the k-mer series from one strand but not the other. To reduce the strand-specific false positives, we implemented a filter requiring identified variants to be supported by reads from both strands, which eliminated all false positives being called by both samtools and nanopolish in the captured region around the TP53 gene. This filter did reduce the number of true positive variants being called by samtools from 11 to 8, but all 14 variants identified by nanopolish were supported by reads from both strands. A visual representation of the detected variants at the *TP53* locus is shown in **Figure 4A**. Raw nanopore reads were error-corrected with the 'phase-reads' module of nanopolish, using the electrical data to interrogate and call SNVs in the single reads at variant locations identified by nanopolish.

Expanding our pipeline to the 8 loci without SVs in GM12878, we called SNVs over a total enriched area of 140kb, where 176 annotated SNVs exist. On the default settings, samtools correctly identified 130 variants (Sensitivity: 0.74, PPV: 0.74) and nanopolish identified 149 (Sensitivity: 0.85, PPV: 0.54). We evaluated the performance of nanopolish variant calling over a range of thresholds, and found the dual-strand-filter performed best when a variant was required to be supported by at least 15% of reads(**Table 2**). At this threshold 134 of the 176 variants were called correctly (Sensitivity: 0.76) with

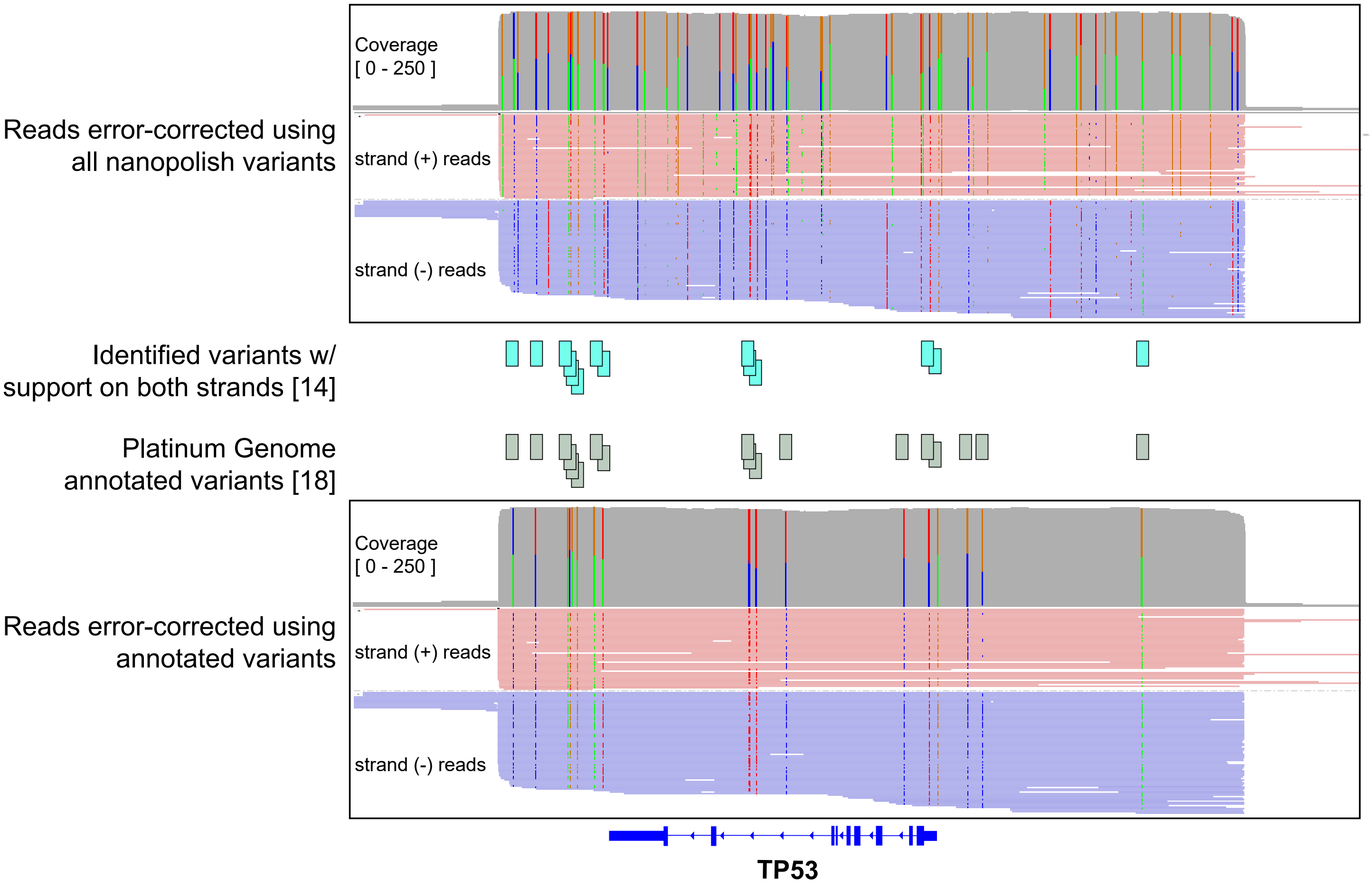


Figure 4 : SNV detection IGV plot of the captured region around the TP53 gene. Reads error-corrected with the 'phase-reads' nanopolish module, using either all variants called by nanopolish (top) or variants from the platinum genome reference9 (bottom). Blue boxes represent nanopolish-called variants passing the dual-strand filter; grey boxes indicate annotated platinum annotated SNVs between guideRNA cut sites.

only one false positive (PPV: 0.99). Closer examination of the few persistent false positives passing the dual-strand filter showed they occur in low-complexity palindromic regions, where the basecaller makes the same mistake on k-mers from both strands (**Supplementary Figure S6b**). Because of the low false positive rate of this method, we delineate variants identified by the dual-strand-filter as our set of 'high-confidence variants' (**Supplementary Table S4a**).

Although the dual-strand filter led to a reduction of false positives, there were instances where the base-caller had difficulty identifying real variants due to errors in base-calling on one of the strands. In an effort to identify variants missed by the dual-strand filter, we utilized the nanopolish 'variant quality score' to build a series of ROC curves (**Figure 4B**) (variant quality score is generated by nanopolish variant calling algorithm and reflects the confidence of the nanopolish HMM). From these ROCs, we found the greatest performance while setting the minimum variant threshold to 25% and requiring variant quality scores to be at least 100. This led to the calling of an additional 13 of the 42 real variants missed by the dual-strand filter, although adding an additional 27 false positive variants (**Supplementary Table S4b**). Because of the higher false positive rate without the dual-strand filter, we designate the additional variants found by this method (and not in the high-confidence list) as 'low-confidence variants'.

In the MDA-MB-231 cell line, we selected 3 annotated single nucleotide variants[22] in genes causally implicated in cancer(*BRAF*, *KRAS*, and *TP53*). Across these 3 regions we called 34 'high-confidence' SNVs in MDA-MB-231, and 22 'low-confidence' SNVs (**Supplementary Table S5a and S5b**). For comparison, the GM12878 cell line has 40 annotated SNVs in these three regions, but in GM12878 variants are restricted to introns and UTRs, unlike in MDA-MB-231 where mutations exist in exons and affect protein coding. The known mutations in BRAF and TP53, were both included in the high-confidence variants, and correctly called as heterozygous and homozygous, respectively. The third MDA-MB-231 missense mutation, in *KRAS*, was found in our 'low-confidence' variants. Systematic errors of the basecaller on the negative strand kept this variant out of the 'high-confidence' set, as it was removed by the dual-strand filter. This shows that the nCATS enrichment method provides a viable strategy to identify single nucleotide variants *de novo* from nanopore signal and it can be used to phase and visualize known variants. Still, there are persisting limitations of nanopore variant calling at present, as evinced by some variants not being detected by this approach.

Discussion

We have described a new method that can be rapidly adapted to evaluate any number of genomic sites simultaneously with high coverage nanopore sequencing data. Because of the low cost to entry and small footprint of the instrument, this assay has the potential to be widely utilized as a tool for evaluating DNA methylation, structural variation and even small mutations.

Our efforts using this method to evaluate single nucleotide variants shows that regions of interest can be queried with the nCATS protocol. As basecaller algorithms continue to improve we anticipate even higher future performance of this tool for the surveillance and identification of mutations. We highlight use of the nCATS method to detect and reassess structural variants. It is only recently, with the advent of long-read sequencing that the great diversity of structural variation in genomes has been appreciated [23,24], and this method provides a dynamic tool to evaluate genomic rearrangements, which are known to contribute to cancer pathogenesis [25]. We show that the assay can be used to evaluate the ploidy of SVs and segregate reads into their parental allele. Importantly, because nanopore sequencing interrogates the DNA strand rather than sequencing-by-synthesis, we can *simultaneously* profile methylation in these loci, providing biological as well as diagnostic insight into the epigenome, which is commonly disrupted in human neoplasia[26]. By positioning guide RNAs adjacent to low complexity regions, this technique can also evaluate DNA methylation in regions of poor bisulfite sequencing mappability[27] while avoiding the high costs associated with whole genome bisulfite sequencing. Thus, this assay represents a fast and inexpensive way to assess clinically relevant genes to detect genomic or epigenomic variation.

The strategies presented here also offer features for additional applications including locating gene insertions from intentional genetic engineering or viral integration, evaluating tandem repeats present in pathological regions, and generating long-read data for building genome assembly scaffolds. By providing a targeted approach that can be rapidly applied to studying DNA at numerous regions of interest, this method helps to expand the applications of nanopore sequencing as an inexpensive alternative for generating high coverage targeted data while maintaining the advantages of long-read sequencing through directly probing the native cellular DNA. By incorporating multiplexing, or using the newly released "flongle" adaptor with cheaper flowcells, this cost may be reduced even further.

METHODS

Cell culture and DNA prep

Cell lines were obtained from ATCC and cultured according to standard protocols. DNA was extracted using the MasterPure kit (Lucigen, MC85200), and stored at 4C until use. DNA was quantified using the Qubit fluorometer (Thermo) immediately before performing the assay.

Guide RNA design

Guide RNAs were assembled as a duplex from synthetic crRNAs (IDT, custom designed) and tracrRNAs (IDT, 1072532). Sequences are provided in **Supplementary Table S1**. The crRNAs were designed using IDTs design tool, and selected for highest predicted on-target performance with minimal off-target activity. The gRNA duplex was designed to introduce cuts on complementary strands flanking the region of interest. For methylation studies and SNV studies, the target size between gRNAs was 20-30 kb; for deletions the gRNAs were designed to flank the suspected breakpoints by ~5kb.

Ribonucleoprotein Complex Assembly

Prior to guide RNA assembly, all crRNAs were pooled into an equimolar mix, with a total concentration of 100uM. The crRNA mix and tracrRNA were then combined such that the tracrRNA concentration and total crRNA concentration were both 10uM. The gRNA duplexes were formed by denaturation for 5 minutes at 95C, then allowed to cool to room temp for 5 minutes on benchtop. Ribonucleoprotein complexes (RNPs) were constructed by combining 10pmol of gRNA duplexes with 10pmol of HiFi Cas9 Nuclease V3 (IDT, 1081060) in 1X CutSmart Buffer (NEB, B7204), allowed 20 min at room temp, then stored at 4C until use, up to 2 days.

Cas9 Cleavage and Library Prep

3ug of input DNA was resuspended in 30uL of 1X CutSmart buffer (NEB, B7204), and dephosphorylated with 3uL of Quick CIP enzyme (NEB, M0508) for 10 min at 37C, followed by heating for 2 minutes at 80C for CIP enzyme inactivation. After allowing sample to return to room temp, 10uL of the pre-assembled 10pmol Cas9/gRNA complex were added to sample. In the same tube, 1uL of 10mM dATP (Zymo, D1005) and 1uL of Taq DNA polymerase (NEB, M0267) were added for A-tailing of DNA ends. The sample was then incubated at 37C for 20min for Cas9 cleavage followed by 5 minutes at 72C for A-tailing. Sequencing adaptors and ligation buffer from the Oxford Nanopore Ligation Sequencing Kit (ONT, LSK109) were ligated to DNA ends using Quick Ligase (NEB, M2200) for 10 min at room temp. The sample was cleaned up using 0.3X Ampure XP beads (Beckman Coulter, A63881), washing twice on a magnetic rack with the long-fragment buffer (ONT, LSK109) before eluting in 15uL of elution buffer (ONT, LSK109). Sequencing libraries were prepared by adding the following to the eluate: 25uL sequencing buffer (ONT, LSK109), 9.5uL loading beads (ONT, LSK109), and 0.5uL sequencing tether (ONT, LSK109).

Sequencing

Each sample was run on a 9.4.1 version flow cell using the GridION sequencer. Initial flow cell priming was performed with 800uL flush buffer (ONT, LSK109), and allowed 5 minutes to equilibrate. The flow cells were then primed with 200uL of a 1-to-2 dilution of sequencing buffer (ONT, LSK109) prior to loading sequencing libraries.

Analysis

Basecalling was performed using the GUPPY algorithm (Version 2.3.7) to generate FASTQ sequencing reads from electrical data. Reads were aligned to the human reference genome (Hg38) using either NGMLR[20] (for SV calling) or Minimap2[28]. Per-nucleotide coverage was determined using samtools, and clustered using the 'bincov' script of the SURVIVOR software package[29].

CpG methylation calling on nanopore data was performed using nanopolish[7]. Methylation calling on existing WGBS data of GM12878 (GEO: GSE86765)[14] was performed using the bismark software tool[30]. RNA-seq data of MCF-10A, MCF-7, and MDA-MB-231 were downloaded from GEO (Accession: GSE75168) in the form of RNA counts.

Deletions were called using the structural variant caller Sniffles[20], set to find deletions with a minimum size of 100bp. Because of the allelic size bias on the very large deletions in GM12878, the ploidy was initially incorrectly called as homozygous. To correct this, we used the option "--min_homo_af" set to 99.9, which ensured a deletion was called as heterozygous if reads supporting the reference were present at a minimum threshold of one in one thousand.

Segregation of GM12878 reads into parental alleles was performed using HapCUT2[19] with existing phased variant information from sequencing of familial trios[10]. For assignment into parental allele, we required at least 75% of variants per read to agree on allele of origin. If variants were not detected or if the variants within a read disagreed on the parental allele it was excluded from allelic analysis.

De novo variant calling was performed using nanopolish[7] or samtools[21]. For calling true positives, we used the platinum genome dataset[10], filtered to contain only SNVs. For calling specificity, the total number of tests were taken as all regions queried with the most permissive nanopolish settings (minimum freq: 10%, minimum quality score: 0).

Accession

Cas9-Enrichment data are available at NCBI Bioproject ID PRJNA531320 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA531320). Code used in data analysis is available at https://github.com/timplab/Cas9Enrichment

Disclosures

This work was supported by funding from NIH R01 HG009190 (NHGRI). JG, ER, RB and AH are employees of ONT. WT has two patents licensed to ONT (US20110226623 A1 and US20120040342 A1). TG, IL and WT have received travel funds to speak at symposia organized by Oxford Nanopore Technologies.

References

1. Isac Lee, Rachael Workman, Josh Zhiyong Wang, Winston Timp. Use of Agilent SureSelect to perform targeted long-read nanopore sequencing. *Agilent Application Note* (04/2017).
2. Gabrieli, T., Sharim, H., Michaeli, Y. & Ebenstein, Y. Cas9-Assisted Targeting of CHromosome segments (CATCH) for targeted nanopore sequencing and optical genome mapping. *bioRxiv* 110163 (2017). doi:10.1101/110163
3. Gießelmann, P. *et al.* Repeat expansion and methylation state analysis with nanopore sequencing. *bioRxiv* 480285 (2018). doi:10.1101/480285
4. Ebbert, M. T. W. *et al.* Long-read sequencing across the C9orf72'GGGGCC'repeat expansion: implications for genetic discovery efforts in human disease. *bioRxiv* 176651 (2018).
5. Leija-Salazar, M. *et al.* Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *Mol Genet Genomic Med* **7**, e564 (2019).
6. Karamitros, T. & Magiorkinis, G. Multiplexed Targeted Sequencing for Oxford Nanopore MinION: A Detailed Library Preparation Procedure. *Methods Mol. Biol.* **1712**, 43–51 (2018).
7. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
8. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *bioRxiv* 504993 (2018). doi:10.1101/504993
9. Messier, T. L. *et al.* Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. *Oncotarget* **7**, 5094–5109 (2016).
10. Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* (2016). doi:10.1101/gr.210500.116
11. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
12. Satya-Prakash, K. L., Pathak, S., Hsu, T. C., Olivé, M. & Cailleau, R. Cytogenetic analysis on eight human breast tumor cell lines: high frequencies of 1q, 11q and HeLa-like marker chromosomes. *Cancer Genet. Cytogenet.* **3**, 61–73 (1981).
13. Rondón-Lagos, M. *et al.* Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis. *Mol. Cytogenet.* **7**, 8 (2014).

14. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

15. Martignano, F. *et al.* GSTP1 Methylation and Protein Expression in Prostate Cancer: Diagnostic Implications. *Dis. Markers* **2016**, 4358292 (2016).

16. Wang, X.-M., Zhang, Z., Pan, L.-H., Cao, X.-C. & Xiao, C. KRT19 and CEACAM5 mRNA-marked circulated tumor cells indicate unfavorable prognosis of breast cancer patients. *Breast Cancer Res. Treat.* (2018). doi:10.1007/s10549-018-05069-9

17. Kabir, N. N., Rönnstrand, L. & Kazi, J. U. Keratin 19 expression correlates with poor prognosis in breast cancer. *Mol. Biol. Rep.* **41**, 7729–7735 (2014).

18. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**, 160025 (2016).

19. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).

20. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

21. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

22. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

23. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* 193144 (2018). doi:10.1101/193144

24. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019).

25. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).

26. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* **13**, 497–510 (2013).

27. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120–e120 (2018).

28. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

29. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).

30. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).